Why Experiment? The Rationale and History of Social Experiments

SOCIAL EXPERIMENTATION:
EVALUATING PUBLIC PROGRAMS WITH EXPERIMENTAL METHODS PART 1

In 1966, a graduate student in economics at the Massachusetts Institute of Technology approached the Office of Economic Opportunity (OEO), President Lyndon Johnson's antipoverty agency, with an unusual proposal. Heather Ross suggested that OEO test the negative income tax (NIT) concept that was then being discussed among academic economists by actually giving money to working poor families and monitoring their behavior.[1] The test would focus on whether the NIT would cause poor families to quit working, as its critics alleged.

What made her proposal different from the many demonstration projects that were funded by OEO as part of the War on Poverty was the suggestion that the test be structured as a classical statistical experiment, with random assignment of families to a "treatment" group, which would be eligible to receive NIT payments, or a "control" group, which would not. The difference in outcomes between the two groups would provide a measure of the effect of the NIT program.

The project that ultimately resulted from Ross's proposal—the New Jersey Income Maintenance Experiment—is generally regarded as the first large-scale "social experiment".[2] Over the period 1968–69, 1,300 low-income families in five cities were randomly assigned to treatment or control status. The effects of the experimental program on family members' employment and earnings, educational attainment, marital stability, and other behavioral outcomes were measured by the difference in subsequent outcomes for the two groups.

As we will see later in this paper, the New Jersey Experiment was not entirely unprecedented. Whether or not it was the first social experiment, though, it certainly sparked widespread interest in the application of experimental methods to a broad range of public programs and policies that had rarely, if ever, been subjected to such rigorous evaluation techniques.

This is the first in a series of papers on the design, implementation, and analysis of social experiments. In this paper, we begin by discussing the rationale for using experimental methods to evaluate public programs. We then examine the ethical issues that are sometimes raised with regard to experiments. The paper concludes with a brief review of the history of the use of experimental methods for program evaluation, both before and since the New Jersey Experiment.

Why Experiment?

In popular parlance, implementing a program on an experimental basis is usually taken to mean "trying out" a new program on a small scale, to see if it "works." If the program is truly new, then only by setting up a pilot test can we obtain any empirical evidence of its effects. So one rationale for social experiments is to create a working model of the program so that we can evaluate it. But determining whether the program "works" requires much more than simply observing it in action. To see why, we must consider what we mean when we say that a program does or does not "work" and what the alternative means of making that determination are. As we will see, the evaluation of ongoing programs raises much the same issues as the testing of new programs.

We take the question of whether a program "works" to mean whether it is effective in achieving its goals. For most social programs, those goals have to do with the effects of the program on its participants. The goal of a job training program, for example, is to raise participants' earnings. If the

[1] Ross (1966).

[2] The formal name of the project was the Graduated Work Incentives Experiment. It was originally implemented in four cities in New Jersey—hence the popular name "New Jersey Experiment"—and subsequently expanded to include families in Scranton, Pennsylvania. See Watts and Rees (1977) for a complete description of the project.

program is targeted on welfare recipients, it might have the further objective of getting them off welfare, or at least reducing their welfare benefits. A teen parenting program might have the goal of inducing teen parents to remain in or return to school, making them better parents, and/or reducing their future child-bearing. A remedial education program might have the objective of raising students' math skills or reading comprehension. We term these behaviors and circumstances of the participants after they enter the program outcomes.[3] To determine whether the program is "working," we must determine whether it was successful in changing the outcomes it is intended to affect.

One can, of course, observe the outcomes of interest—employment, receipt of welfare, school attendance, etc.—simply by following participants after they leave the program. But one cannot, on the basis of participant behavior alone, know what portion of the observed post-program outcomes should be attributed to the program. We define the impact of the program as the difference between the observed outcomes of participants (*e.g.*, their post-program earnings) and *what those outcomes would have been in the absence of the program*. Measuring the actual outcomes of participants is relatively straight-forward; not surprisingly, measuring what they would have done in the absence of the program is much more difficult.

Suppose, for example, that 80 percent of the graduates of a job training program obtain jobs when they leave the program. Does that mean that the program is achieving its objective of increasing the employment and earnings of its participants? Not necessarily. Some of the participants would have gotten jobs even if they hadn't gone through the training program. If only 20 percent would have found jobs in the absence of the program, then the program has increased participants' employment rate by 60 percentage points. But if 80 percent of the participants would have gotten jobs without the program's help, the program has had zero impact on participants' employment rate. (Of course, in that case, it may be helping them get *better* jobs, or get jobs faster, so we would need to look at outcomes like earnings and hours of work, as well as employment rates.)

Similar reasoning applies to most social programs. Some individuals would have gotten off welfare, or returned to school, or improved their educational performance even without special assistance. We can only attribute behavioral changes over and above that base level to the program. *The fundamental problem of program evaluation, then, is to determine what would have happened in the absence of the program*.

There are a number of different ways to attempt to measure what would have happened in the absence of the program. To understand why experimental methods are the preferred approach, it is useful to understand the shortcomings of other approaches. We therefore begin by discussing the two principal alternatives to experimental designs: pre–post designs (sometimes called "reflexive" designs) and comparison group designs. Our discussion, here and elsewhere in this series, focuses on programs where the principal effects of interest to the evaluator are on individual program participants, rather than on institutions or the broader community. The discussion encompasses both the evaluation of ongoing programs and the testing of new programs in special demonstrations.

Pre–Post Designs

A particularly simple way to attempt to determine what would have happened in the absence of the program is to use the behavior of the participants before they came into the program. In the case of a job training program, for example, we might use the earnings of the participants in the year before they applied to the program. Or, in a remedial education program for high school students, we might use the students' grades in the previous school year. We would then measure the impact of the program by the change in earnings or grades between the year prior to program participation and the year after. This design has the advantage of requiring only data on participants.

Pre–post designs are not always feasible. The institutional setting may preclude collection of the relevant data before the intervention begins. Or the outcome of interest may not be defined in the pre-program period. Suppose, for example, that we are evaluating a program of prenatal care and the principal outcomes of interest are measures of the health of the baby; these measures are not defined in the pre-program period.

Even where pre-program data can be collected, though, behavior in the period prior to program entry may not be a good predictor of what would have happened later on in the absence of the program, for several reasons. Factors external to the program may change over the same time period. In the case of the training program, for example, an improvement in the local economy might result in increased earnings, quite aside from any effect of the program. A pre–post impact measure would erroneously attribute this improvement in earnings to the program.

[3] Because some program effects can occur almost immediately, while the participant is still in the program, we define outcomes to include everything that happens after program entry, not just what happens after the participant leaves the program.

Participants' outcomes may also change over time because of natural maturation processes. Suppose, for example, that we wish to measure the effect of a pre-school program on the social skills of children. Even without a special program, children's social skills can be expected to improve over time. Thus, a pre-post impact measure would overstate the effects of the program by including gains in social skills that would have occurred anyway. Much the same type of maturation effects are likely to affect the employment and earnings of teenagers entering the job market; pre-post measures are therefore likely to overstate the impact of job training programs on young workers.

Even when there are no pronounced secular trends in the outcome measure for the population at large or maturation effects, pre-post designs will yield misleading conclusions if, on average, the pre-program period was an atypical one for participants. For example, individuals usually apply to job training programs when they are unemployed. Even without the program's assistance, though, some of the participants would eventually find work on their own. Thus, in the absence of the program, job training participants would show a rising earnings trend from the period before program entry to the period after. Statisticians refer to this as "regression to the mean."

This phenomenon of individuals participating in public programs when their outcomes of interest are atypically low (or high) is not confined to job training programs. People naturally tend to apply to social programs when their need is greatest. Moreover, social programs often select participants on the basis of need—as measured by the same indicators that evaluators use to measure the impact of the program. To the extent that these needs would be only temporary even without the program's assistance, selection on need will result in regression to the mean—with need greatest around the time of program entry, then declining over time. In such cases, simple pre-post differences in outcomes will overstate the effect of the program by including the rebound from this temporary need for assistance, which would have occurred even without the program's help.

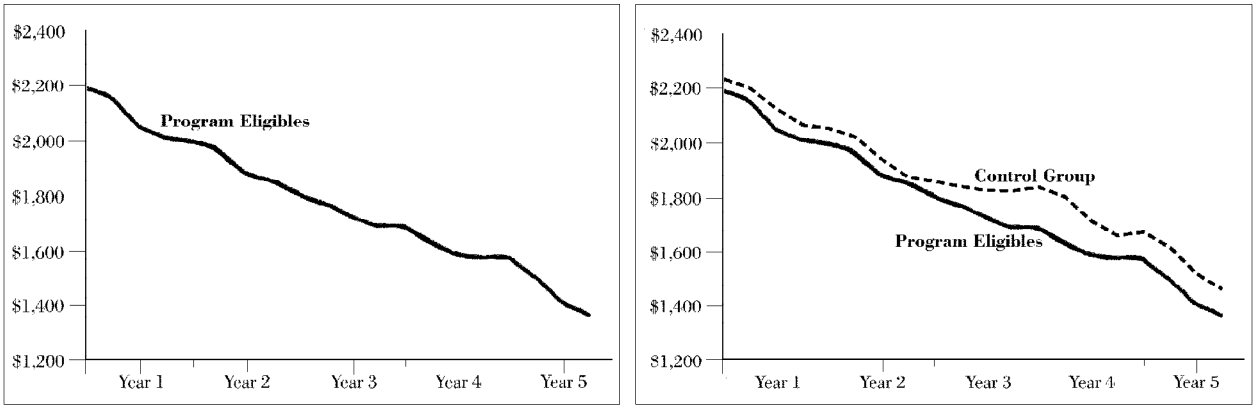
Exhibit 1 shows how failure to take into account the temporary nature of the need for assistance can create a misleading impression of program effectiveness. The left hand panel of the exhibit shows the time path of quarterly public assistance benefits to a sample of AFDC recipients who were eligible for a program designed to help them become employed and leave the welfare rolls. The steep downward trend in assistance payments would appear to indicate that the program was quite effective. This is in fact the type of "evidence" that is frequently used to demonstrate program effectiveness in the legislative process and in the popular media. As shown in the panel on the right, however, the decline in benefits was nearly as steep for a control group of program eligibles who were *excluded* from the program, as part of a social experiment. This comparison demonstrates that most of the decline in benefits experienced by the participant group was the result of normal turnover of the welfare rolls, as recipients' circumstances improved and they were able to leave welfare. Attributing the effects of this turnover to the program greatly overstates its effectiveness.

Comparison Group Designs

As the previous example suggests, one way to avoid some of these hazards of pre-post designs is through the use of a

Average Quarterly Assistance Payments, New York Child Assistance Program

EXHIBIT 1



Illustrative Comparison Group Design

EXHIBIT 2

Group	Pre-program GPA	Post-program GPA	Pre-Post Change
1. Participants	2.0	2.6	0.6
2. Comparison Group	2.2	2.4	0.2
3. Estimated Impact			0.4

comparison group. This approach involves selecting a group of individuals who are as similar as possible to the participants, except that they do not participate in the program, and monitoring their outcomes. The pre–post change in outcomes of the comparison group is used to represent what would have happened to the participants in the absence of the program. Ideally, the pre–post change in earnings of the comparison group will reflect any general rise in earnings in the local labor market, natural maturation factors, or rebound from a spell of unemployment. In the simplest case, then, subtracting the pre–post change in outcomes of the comparison group from the pre-post change in outcomes of the participants should yield a more accurate measure of the impact of the program.

Exhibit 2 illustrates this approach for a hypothetical remedial education program. The outcome of interest, student grade point average (GPA), was measured in the semester before and the semester after the program, for participants and for a comparison group of nonparticipants. The first row of the exhibit shows that a simple pre–post measure of impact, using data for participants only, would indicate that the program increased GPA by 0.6 points. In the comparison group, however, GPA rose 0.2 points even without the assistance of the program (see row 2). Under the assumption that participants would have experienced that same rise in GPA in the absence of the program, the comparison group approach estimates the impact of the program as .4 points—the change in GPA from pre- to post-program for the participant group less the pre-post change for the comparison group (see row 3).

The comparison group design thus avoids the erroneous attribution of the full increase in GPA to the program that would occur with a simple pre-post measure of impact. It is, however, still potentially subject to bias associated with the way the treatment group and comparison group were selected. Because the pervasiveness of “selection bias” in nonexperimental comparison group designs is one of the most important reasons for preferring experimental methods, we now examine this phenomenon in some detail.

Selection Bias

The comparison group design is based on the assumption that program participants would have experienced the same change in outcomes as the comparison group had they not gone through the program. Estimates based on comparison group designs are only as valid as that assumption. And ultimately, one can never be sure how valid that assumption is because it is a statement about something that is inherently unobservable—the experience of the program participants if they had not entered the program.

What we do know is that the participants were either self-selected or selected by somebody else (*e.g.*, a teacher or a welfare case worker) to go into the program, whereas the comparison group members were not. Unless those selection decisions were totally random, this means that the two groups differ in some way. If the difference(s) that led one group to be selected for the program and the other not to be selected also lead to differences in the outcomes of interest, the comparison group design will erroneously attribute those differences in outcomes to the impact of the program. Such errors in attribution are termed “selection bias.”

Suppose, for example, that the hypothetical remedial education program discussed above was open, on a voluntary basis, to all students with GPAs below 2.5 and that the comparison group was composed of all students who were eligible but did not volunteer to participate. The fact that the participants volunteered for the program may suggest that they are more motivated than the students in the comparison group, who did not—and that may suggest that their improvement in grades would have been greater than that of the comparison group even without the program’s help.

Exhibit 3 illustrates how this would lead to bias in the estimate of program impact. The first three lines of the exhibit simply reproduce the information in Exhibit 2. The fourth line shows the (unobservable) pre-post change in participants’ GPA that would have occurred in the absence of the program—here assumed to be more than that of the comparison group because the participants are assumed to be more highly motivated. The true impact of the program is

Illustrative Comparison Group Design, Relative to True Impact

EXHIBIT 3

Group	Pre-program GPA	Post-program GPA	Pre-Post Change
1. Participants	2.0	2.6	0.6
2. Comparison Group	2.2	2.4	0.2
3. Estimated Impact			0.4
4. Participants, without program	2.0	2.5	0.5
5. True Impact			0.1

the difference between the observed pre-post change in participants' GPA less the change that would have occurred in the absence of the program. As shown on line 5, in this case the true impact is 0.1. The estimate based on the comparison group (0.4) overstates the true impact by the difference between the pre-post change in the comparison group and the true without-program pre-post change for the participants.[4]

Alternatively, suppose the participants were selected by their teachers. If the teachers selected those who were least likely to do well without the program's assistance, the comparison group's grade gains might well overstate the change in grades that could be expected for participants in the absence of the program. In that case, the estimate based on the comparison group would understate the effect of the program—*i.e.*, it would be biased downward.

Selection bias encompasses any differences between the program participants and the comparison group that affect the outcomes of interest. Suppose, for example, that a comparison group for the participants in a job training program is selected from communities where the program is not conducted. Differences between the labor markets in the program communities and the comparison communities may cause the employment and earnings of the comparison group either to overstate or to understate what would have happened to the program participants in the absence of the program.

One can, of course, attempt to match the comparison group to the participants in terms of such personal characteristics as age, race, gender, prior employment experience or grades (depending on the nature of the experiment), and/or environmental characteristics such as the local unemployment rate or rural vs. urban setting. Comparison groups are sometimes drawn from national survey data bases like the Current Population Survey or the decennial Census, using such matching techniques. But one can only match on measured characteristics. If the two groups differ in unmeasured characteristics, such as motivation or native ability, their outcomes may differ for reasons that have nothing to do with the program.

Experimental Designs

As noted above, the central problem in measuring the impact of a program is that we cannot observe what the participants' outcomes would have been in the absence of the program. We can try to represent those outcomes with those of a comparison group, but if there are systematic differences between the comparison group and the participants that affect the outcomes of interest, impact estimates based on the comparison group will be biased.

Random assignment offers a way to create a comparison group that is not systematically different from the participants—*i.e.*, one that is not subject to selection bias. If assignment to the program or to the comparison group is completely random, selection into one group or the other is *by definition* unrelated to any characteristic of the individual—and therefore to the individual's subsequent outcomes. Thus, any systematic differences in post-random assignment outcomes between the two groups can confidently be attributed to the experimental program.[5]

[4] In this and other examples presented in this section, for simplicity and clarity of exposition we abstract from the sampling variability of the estimates. Determining empirically whether a particular impact estimate is biased is an extremely complex matter once sampling variability is taken into account (see Bell et al., 1995, for a discussion of this issue).

[5] By "systematic" differences, we mean any differences that are larger than might be expected on the basis of sampling error. In a later paper, we will discuss how one can test whether the treatment-control difference could be due to sampling error.

The defining characteristic of a social experiment is random assignment of some pool of individuals to two or more groups that are subject to different policy regimes. One of these groups is a control group that is subject to the existing policy environment—*i.e.*, it is excluded from the experimental program.[6,7] In addition, one or more treatment groups consist of individuals assigned to one or more variants of the program being evaluated. Data on the relevant outcomes of each group are then measured over some follow-up period and the impact of the program is estimated as the difference between treatment and control group outcomes.

By "random assignment" we mean assignment of individuals to groups on the basis of a random event, such that each individual has a specified probability of being assigned to each group.[8] The random event can be as simple as the flip of a coin—if the coin comes up heads, the individual is assigned to the treatment group; tails, he or she is assigned to the control group. In this case, each individual would have a 50 percent chance of assignment to each group and, if large numbers of applicants are randomly assigned, the total sample will be divided approximately evenly between the two groups. In practice, random assignment is usually based on specially designed tables of random numbers or computer algorithms that generate random numbers.[9] Whatever method is employed, the important thing is that each individual have a specified probability of being assigned to each group and that the assignment itself be made by chance alone.

Suppose, for example, that we wish to evaluate a new job training program. We would begin by setting the program up on a pilot basis and recruiting applicants. Those judged eligible (and still interested after learning more about the program) would then be randomly assigned to a treatment group, which is allowed to enter the program, or to a control group, which is not. Since the primary objective of a training program is to raise participants' earnings, we would collect data on the earnings of individuals in both groups following random assignment. The impact of the program on participants' earnings would be measured by the difference in mean earnings between the treatment and control groups over the follow-up period. Because the two groups were well-matched at the point of entry into the program, it is not essential to compare the *changes* in the two groups' earnings over the follow-up period, as we did in the comparison group design. The simple treatment-control difference in post-random assignment earnings will, on average, give the same answer as the treatment-control difference in *changes* in earnings, because there was no systematic difference between the two groups at the point of random assignment. (As we will see in a later paper, however, taking the pre-program value of the outcome variable into account will improve the precision of the impact estimates.)

While this simple example illustrates the fundamental elements of a social experiment, the range of possible variations on this basic theme is enormous. For example, instead of evaluating a single program, one might wish to compare several different program models or estimate the effects of specific program components. Instead of studying the effects of the program on applicants who voluntarily apply to the program, one might wish to study its effects on the entire population eligible for the program, or some subset of eligibles. The program to be evaluated need not be a new one; it could be an ongoing program. And policymakers may be interested in the effects of the program on a whole range of participant outcomes, including some that are difficult to quantify, not just a single, easily quantified outcome like earnings. In subsequent papers in this series, we will explore the design variations that will allow the experimenter to address these and other evaluation objectives.

Experimental vs. Nonexperimental Impact Estimates—An Empirical Example

The difference between nonexperimental comparison groups and randomly assigned control groups—and therefore the potential for selection bias in impact estimates based on comparison groups—can be quite striking. Exhibit 4 compares the time path of earnings for an experimental control group of job training applicants with that of a comparison group of individuals who were eligible for the program, but did not apply.[10] As shown in the exhibit, controls' earnings fell sharply in the months prior to application to the pro-

[6] Although experimental control groups are a type of comparison group, for the sake of clarity we will generally use the term "comparison group" to mean a nonexperimental comparison group. The term "control" group always denotes random assignment.

[7] In rare instances, where the principal policy interest is in comparison of alternative new policies, the experiment may not include a control group subject to current policy.

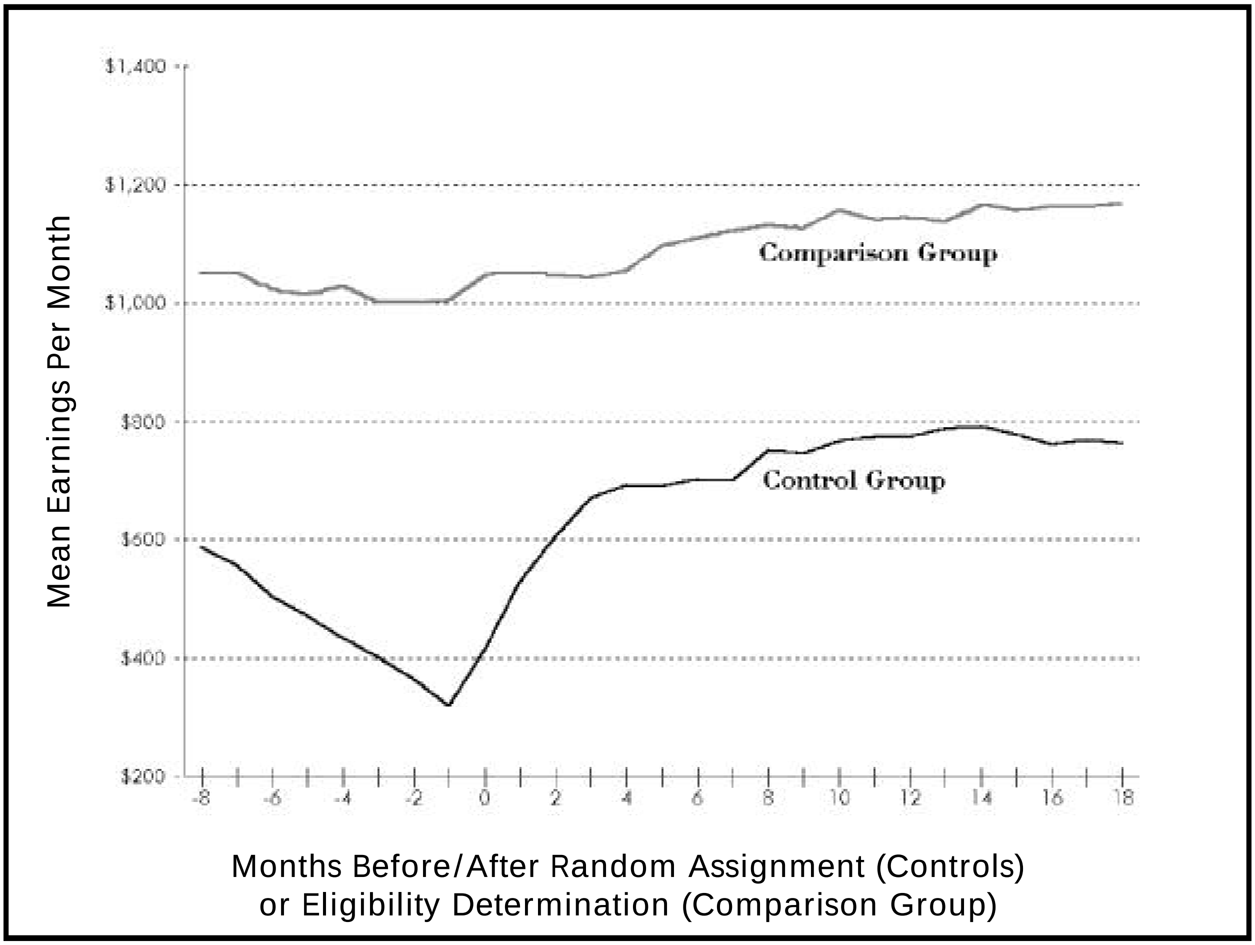
[8] It is important to recognize that "random" assignment does not simply mean haphazard or arbitrary assignment. In practice, great care must be taken to ensure that each individual assigned has the prescribed probability of assignment to each group. We will return to this topic in a later paper.

[9] In a subsequent paper, we discuss procedures for implementing random assignment in some detail.

[10] Exhibit 4 is based on data from the National JTPA Study (unpublished tabulations). The comparison group shown here is composed of a representative sample of JTPA eligibles in four study sites, identified through a screening survey of randomly selected households. Both the control group and the comparison group are composed of adult men.

Mean Earnings, Experimental Control Group vs. Nonexperimental Comparison Group, National JTPA Study

EXHIBIT 4



gram, then rebounded after they applied to JTPA. In contrast, the time path of earnings of the eligible nonparticipant comparison group shows only a slight upward trend over this 2½ year time period.

This stark difference in earnings paths reflects the difference in the two groups' situations at the time they were selected. Controls were selected at a point in time when many of them had just experienced a spell of unemployment, leading them to apply to JTPA.[11] In contrast, the comparison group members were selected solely on the basis of their eligibility for JTPA, without regard to their recent work history. Since JTPA eligibility is determined primarily by family income, this selection procedure ensures that the comparison group members will have relatively low earnings. But since selection into the comparison group was not keyed to any particular event in the individuals' lives, the average earnings of the group is stable over time.

To see how these differences in earnings paths can lead to biased impact estimates, consider Exhibit 5(next page), which presents alternative estimates of program impact, based on the data in Exhibit 4, together with data on the earnings of program participants. As can be seen in Exhibit 5, an impact estimate based on the comparison group of eligible nonapplicants erroneously attributes a large share of the change in participants' earnings from pre-program to post-program to the impact of the program (compare lines 1 and 3) because there was very little change in comparison group earnings over time (see line 2). In contrast, the earnings path of the control group shows that almost all of this change would have occurred even in the absence of the program. The comparison group–based estimate, therefore, substantially overstates the impact of the program.

[11] In the employment and training literature, the decline in earnings prior to program entry is known as the "pre-program dip."

Comparison Group Based Impact Estimate vs. Experimental Estimate of JTPA Impact on Annual Earnings

EXHIBIT 5

Group	Pre-program Earnings	Post-program Earnings	Pre-Post Change
1. Participants	$ 6,300	$ 8,900	+ $ 2,600
2. Comparison Group	12,000	13,200	+ 1,200
3. Estimated Impact			+ 1,400
4. Experimental Control Group	6,300	8,800	+ 2,500
5. Experimental Impact Estimate			+ 100

It must be noted that there are statistical methods that can be used to attempt to adjust nonexperimental impact estimates for differences between the comparison group and the participants.[12] These methods rely, however, on assumptions about the behavior of the two groups that are generally impossible to verify. Thus, in the end, one can never be certain whether the method successfully adjusted for all the relevant differences between the two groups. The multiplicity of nonexperimental estimators available compounds the uncertainty; since different nonexperimental methods generally give different answers, it is unclear which method one should apply.

Detailed discussion of these methods is beyond the scope of this paper. Fortunately, it is also unnecessary. The great methodological advantage of experimental methods over nonexperimental estimators does not lie in the technical details of either approach. It lies in the fact that, unlike any nonexperimental method, properly implemented random assignment guarantees impact estimates that, aside from sampling error, reflect only the effects of the experimental treatment. Nonexperimental methods may be equally reliable in any given application; we simply cannot know *a priori* that they are, as we can with experimental methods.

Internal Validity and External Validity of Experimental Estimates

Evaluation methods that provide unbiased estimates of the impact of *the specific program tested on the population to which it was applied* are termed internally valid. Unlike nonexperimental methods, properly implemented social experiments are guaranteed to provide internally valid impact estimates. To be useful for policy purposes, however, impact estimates must have both internal validity and external validity. Externally valid estimators provide unbiased estimates of the impact of *the program of interest for policy purposes on the population to which it is to be applied*. That is, only if the experiment faithfully replicates the program of interest to policymakers and applies it to a sample that is representative of the policy-relevant population will it have external validity and provide a reliable guide for policy decisions.

The external validity of the experiment can be compromised in a number of ways. For example, because experiments take some time to conduct, the policy of interest often evolves and changes while the experiment is underway, so that when the results become available they represent a somewhat different intervention than the one under consideration at that time. And for reasons of cost and logistics, experimental samples are usually clustered in a small number of localities; this makes it difficult to draw a sample that is truly representative of the entire U.S. population.

In a subsequent paper, we will consider ways to protect against these and other threats to the external validity in the implementation of the experiment. In the end, true external validity is an ideal that is almost impossible to attain, if only because the continually evolving policy process represents such a moving target. Nevertheless, it is an important ideal to strive for, and in assessing the strengths and weaknesses of alternative evaluation methods or results, it is important to gauge their external validity as well as their internal validity.

[12] See Heckman and Robb (1985).

Is Experimentation Ethical?

Because they are the only known way to be confident of avoiding the risk of selection bias, experimental designs have become the preferred method of program evaluation among most of the policy research community and among a growing number of policy makers. But such designs are also subject to a widespread concern, especially among program practitioners: Do they violate ethical standards by denying program services to the control group?

It should be noted at the outset that posing the question this way *presumes* that the program services are beneficial. This is not necessarily the case. Job training programs have been known to *reduce* the earnings of their participants and the offer of wage subsidies to employers as an inducement to hire welfare recipients has been shown to reduce their employment rate.[13] In fact, the very existence of the evaluation is evidence that the agency that funded it is uncertain about the value of the services. Thus, the fact that the control group is denied program services does not automatically mean that they are disadvantaged by the study.

Nevertheless, it is important to know whether controls are in fact "denied" services in any meaningful sense. To answer this question, we must consider several distinct experimental contexts:

- Special demonstration programs set up explicitly to study the effects of a new program;
- Ongoing programs that can only accommodate a limited number of participants; and,
- Ongoing programs that accept all eligible applicants—so-called "entitlement" programs.

Demonstration Programs. There is little disagreement that random assignment to a control group is ethical when it occurs in the context of a special, small-scale demonstration to test a new program. In that context, denial of program services or benefits to the control group simply leaves them in the same position they would have been if the demonstration had never occurred. It is not so much that such demonstrations deny benefits to controls as that they do not provide them to everybody. In this context, random assignment can be viewed simply as a way to ration limited program resources among those who apply to the demonstration. To argue that such a demonstration should not be conducted because it does not provide benefits to everybody who wants or needs them is to argue that programs should never be tested on a pilot basis before full-scale implementation. Clearly, from the standpoint of society as a whole, it is more ethical to conduct a small-scale test of a new program before opening it to the entire target population, because it could have harmful effects or, if not actually harmful, could be an ineffective waste of resources. And if we are to test new programs, we should use the most accurate, reliable evaluation methods available to do so.

Ongoing Programs with Limited Enrollments. Much the same argument applies to ongoing programs that serve less than their entire eligible population. Many programs fall into this category; each year Congress appropriates fixed amounts for job training programs, housing subsidies, and child care assistance that are far less than what would be required to provide services or benefits to all individuals who are eligible for them. Program administrators respond to this shortfall in a number of ways. In some programs, excess demand is rationed by waiting lists, on a first-come, first-served basis or on the basis of service priorities. In other programs, the flow of applications is controlled by varying the amount of program outreach and recruiting activity. Some programs simply turn away those applicants who, in the judgment of program staff, are least likely to benefit from the program.

In this context, random assignment *need not reduce the total number of individuals served by the program*. If there is excess demand for program services or benefits, random assignment may simply reallocate program services or benefits to a *different* set of participants. The ethical issue then becomes, is random assignment a more ethical way to ration scarce resources than the rationing device the program would otherwise have used?

Suppose, for example, that there are 150 applicants to a housing subsidy program, but that the program's budget will only accommodate 100 families. One solution would be to give subsidies to the first 100 families who apply and turn away the remaining 50. Another might be for program staff to exercise their judgment and provide the subsidies to the 100 most "deserving" applicants. Alternatively, one could randomly assign 100 families to receive subsidies and 50 families to a control group. Because it gives each family an equal chance of receiving a subsidy, random assignment is arguably a fairer way to allocate the scarce subsidy funds than either a first-come, first-served policy or staff judgment. In addition, it has the added social benefit of generating knowledge about the effects of the program; this knowledge can then be used to improve the program, to the benefit of these and other similar families.

[13] See Orr et al. (1996) for an example of the former and Burtless (1985) for an example of the latter.

Ongoing entitlement programs. In an ongoing program that provides services or benefits to all eligible individuals (or all that apply), random assignment to a no-service control group *would* constitute denial of services, even in the aggregate. To decide whether an experiment would be ethical in this situation, one would have to weigh the harm done by that denial of service against the social benefits of the knowledge to be gained from the experiment. To date, evaluators and policymakers have taken the position that it is not permissible to deny entitlements for research purposes; this does not mean, however, that there are not instances in which such denial would be justified.

The ethical problems raised by denial of entitlements argue strongly for thorough testing of programs on a small scale *before* they are applied to the population at large. Once a service or benefit has become an entitlement, further testing of its effects becomes extremely problematic, if not impossible. This means that if an entitlement is actually harmful, or simply ineffective, we might never know that and the harm, or waste of resources, might be perpetuated indefinitely.

There are, however, ways in which at least some entitlement programs might be ethically evaluated with experimental methods. One method is to compare the effects of the existing program with those of an alternative program that provides comparable benefits. For example, it might be deemed unethical to deny food stamps to a control group in order to test the effects of food stamps on the nutrition of low-income families. But most observers would agree that it is ethical to *replace* food stamps with their cash equivalent for a randomly assigned control group, in order to measure the nutritional effects of earmarking the subsidy for food. Such an experiment would not, of course, reveal any effect—positive or negative—that are common to both modes of subsidy.

More generally, it is sometimes possible to compensate subjects for any loss of benefits they may suffer as a result of participating in the experiment. In the Health Insurance Experiment, families were asked to give up their existing health insurance policies and accept specially designed policies provided by the study. Some of the experimental policies contained "cost-sharing" provisions that required the family to pay a portion of the cost of the care they consumed, in order to allow the researchers to estimate the effect of the net price of medical care (*i.e.*, its cost to the family) on the use of medical care. In those cases where the family's existing policy covered a larger fraction of the cost of care than the experimental policy, a lump-sum cash payment was made to the family to make up the difference.[14]

As this example suggests, however, in compensating experimental subjects for loss of benefits, one must be careful not to change the experimental treatment or the treatment-control contrast.[15] It might be argued that the lump-sum payments in the Health Insurance Experiment offset the cost-sharing provisions of the experimental plans. However, because they were unrestricted cash payments that could be used for any purpose and were unrelated to the amount of medical care used by the family, the lump-sum payments did not affect the price of medical care to the family; at most, they may have had a small income effect on the family's consumption of care.

Other Ethical Considerations. In considering whether experiments are ethical, it is important not to focus too exclusively on the issue of denial of services to controls. Other members of society have a stake in whether the experiment is performed. In particular, failure to obtain reliable estimates of the efficacy of an ongoing program can entail substantial costs to the taxpayers who support it. An ineffective program can waste millions or billions of the taxpayers' dollars year after year.

Moreover, failure to detect ineffective programs imposes costs on the intended beneficiaries of those programs. Not only do such programs waste participants' time and create false expectations, but they also consume resources that might otherwise be devoted to more effective solutions to the problems those programs were intended to address. Thus, "protecting" program beneficiaries from experiments is not necessarily in their best interest.

Similar considerations apply in the case of demonstrations of new programs. In the absence of reliable knowledge about program effectiveness, ineffective solutions are likely to be legislated, with the same attendant waste of tax resources,

[14] To ensure that no family could be made worse off financially by participating in the experiment, the annual lump-sum payments were set equal to the maximum difference in medical costs that the family could experience during the year.

[15] Alternatively, one can explicitly change the question under study to reflect the actual treatment-control contrast. The earlier example of comparing the nutritional effects of food stamps (in the treatment group) with their cash equivalent (in the control group) can be viewed as a case of compensating controls for loss of benefits. In that case, we reformulated the question under study from, "What are the nutritional effects of food stamps, as compared to no assistance?" to "What are the nutritional effects of earmarking assistance for food, as compared with cash assistance?". In general, however, the question to be studied should be determined by the policy issues that prompted the study, not the feasibility of the experimental design.

disappointed expectations, and displacement of more effective programs.

These considerations do not imply that experiments are always ethically sound. But they do suggest that well-designed experiments addressing important policy issues can have great social value, and that that value must be weighed against any loss to the experimental subjects in deciding whether a particular experiment is ethical.

Informed Consent.[16] One approach that is often suggested, and frequently used, to attempt to protect experimental subjects is to require that experimenters obtain the subjects' informed consent. This involves giving subjects a complete description of the experimental procedures, including any risks to the subject, and obtaining their voluntary consent to participate in the experiment.

Informed consent was developed by medical researchers to protect patients from unwittingly being subjected to experimental medical procedures that might actually be harmful to them. Properly implemented, it is an effective device for this purpose, and we strongly recommend that it be employed in any experiment where the treatment entails any risk of harm to the subject.

More generally, in the context of a demonstration to test a new program, informed consent ensures that each sample member views participation in the experiment as beneficial to him or her. In that case, refusal to consent—and therefore exclusion from the demonstration—leaves the individual no worse off than he or she would have been in the absence of the experiment. Thus, the individual will only consent to participate if, in his or her judgment, the experiment conveys positive net benefits.

In the case of an ongoing program, however, the informed consent of the applicant cannot be taken to mean that he or she expects the experiment to convey net positive benefits *relative to his or her situation in the absence of the experiment*. In that case, the applicant may have received program services in the absence of the experiment. Therefore refusal to consent, resulting in exclusion from the program, leaves the applicant worse off than he or she would have been in the absence of the experiment.[17] Thus, consent implies only that the applicant prefers some chance of receiving program services to no chance at all. To ascertain whether the typical applicant is worse off than he or she would have been in the absence of the experiment, one must determine whether the experiment reduces the total number of applicants accepted into the program, and therefore the probability of acceptance, as discussed earlier.

Even in this case, informed consent does ensure that potential experimental subjects receive a thorough explanation of the experimental treatment and procedures, including any attendant risks. And it is useful to ask the subject to sign a form outlining those procedures, to document that they have been so informed. But it is important to recognize that, unlike the case of a special demonstration, in an ongoing program informed consent does not speak to the issue of denial of services to controls.

A Brief History of Social Experimentation

As noted at the outset, the New Jersey Income Maintenance Experiment of the late 1960s marked the beginning of sustained interest in the use of experimental methods to evaluate social policies. In this section, we briefly review the intellectual history of social experiments, their growing use and widespread acceptance, and the influence they have had on policy.

The Origins of the Experimental Method in the Social Sciences. Social experimentation has a superficial resemblance to the laboratory experiments that have been well-established in the physical and biological sciences for over 200 years. In both cases, the outcomes of different "treatments" are carefully measured and differences in those outcomes are attributed to the difference in treatment.

But social experiments differ from laboratory experiments in one very crucial respect. Laboratory researchers attempt to isolate the effects of treatment by directly controlling the research environment so that the materials or animals to which the alternative treatments are applied, and the conditions under which they are applied, are identical and the only difference lies in the treatment itself. In social programs, direct control of all the factors that might influence the outcomes of interest (*i.e.*, the behavior of the people who make up the sample!) is unattainable. Instead, the social experimenter uses random assignment to ensure the *statistical* equivalence of the different treatment groups—*i.e.*, to ensure that they do not differ *systematically* in ways that could affect the outcomes. The experimenter then applies statistical tests to the outcomes to distinguish the effects of the treatment from the chance variation produced by random assignment.

[16] For a more extended discussion of informed consent in the context of social experiments, see Gramlich and Orr (1975).

[17] In experimental evaluations of ongoing programs, applicants who refuse to consent to random assignment must be excluded from the program. Otherwise, all applicants would have an incentive to refuse to consent, since refusal would increase their chances of getting into the program.

The power of random assignment to eliminate bias by establishing comparable groups was recognized by educational researchers as early as the 1920s. Campbell and Stanley (1963) credit W.A. McCall with having this insight in his 1923 book *How to Experiment in Education*. The great statistician R.A. Fisher laid the statistical foundations of experimentation with random assignment in his seminal books *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935).

Since the 1930s, random assignment has been used routinely in educational and psychological research, usually with small groups of students exposed to different teaching methods or psychological stimuli. Campbell and Stanley's own 1963 classic, *Experimental and Quasi-Experimental Designs for Research*, clearly laid out all of the issues discussed so far in this paper, in the context of educational research. Over the same period, random assignment became one of the dominant modes of medical research, with patients randomly assigned to receive experimental drugs or medical procedures, for comparison with control groups receiving standard treatments or placebos.

Application of Experimental Methods to Social Programs. By the 1960s, the use of the experimental model was sufficiently widespread in education, psychology, and medicine that it was quite natural to apply it to social programs and policies outside those fields. In 1961-64, for example, the Manhattan Bail Bond project used random assignment to test the proposition that many individuals could successfully be released without bail prior to trial.[18] Other experimental tests of law-related programs and procedures undertaken in the 1960s included studies of a variety of approaches to the prevention and treatment of juvenile delinquency, the use of legal counsel in juvenile court, the effects of pretrial conferences, low-stress vs. high-stress training for police, alternative penalties for drunk driving, and vocational, surgical, and social rehabilitation for former prisoners.[19]

These early applications to social policy received little attention outside the immediate circles of the researchers and funding agencies involved, however. Therefore, in 1967 when the proposal was made to use random assignment to evaluate the NIT concept, it was viewed as a totally novel idea. And in many ways it was. The New Jersey Income Maintenance Experiment marked the first use of experimental methods to test a proposed social policy in the field on a large scale. The sample of 1,300 families randomly assigned in the New Jersey Experiment was larger than the samples in most of the experiments that had come before. More importantly, the experiment involved administering carefully controlled treatments to, and observing the behavior of, this large sample of individuals in the course of their daily lives, not in a classroom, hospital, or other institutional setting. In addition, the question addressed by the experiment—whether receipt of welfare would cause poor families to stop working—was a highly visible, politically charged issue.

The New Jersey Experiment represented the marriage of the statistical tradition described above with the demonstration programs that flourished in the Great Society era of the mid-1960s.[20] Funded by a number of Federal agencies, but most notably by the new antipoverty agency, the Office of Economic Opportunity (OEO), these demonstrations were intended as pilot tests of service delivery models that their designers hoped would ultimately be implemented on a national scale. The typical demonstration program was designed more to mobilize political support for the program than to measure its effects; few involved careful data collection and, prior to 1967, none involved a rigorous research design. The New Jersey Experiment imposed statistical rigor on this normally chaotic field enterprise.

The New Jersey Experiment was designed to address an issue that had stymied advocates of the negative income tax concept: would cash transfers to the working poor cause them to substantially reduce their work effort, as critics of the policy alleged? Previous efforts to address this question with existing data had yielded very inconclusive results. Because large-scale cash transfers to this population had never been implemented before, nonexperimental studies of this question, using survey data on the national population of working poor families, essentially compared the labor supply of low-income individuals who received such "unearned income" as unemployment compensation, veterans' benefits, and workman's compensation with that of individuals with no such income. Such comparisons are subject to severe selection bias, since eligibility for these forms of income is determined in part by the individual's past and present work effort. As a result, different nonexperimental methods yielded widely varying estimates. The experiment was intended to resolve this crucial political issue.

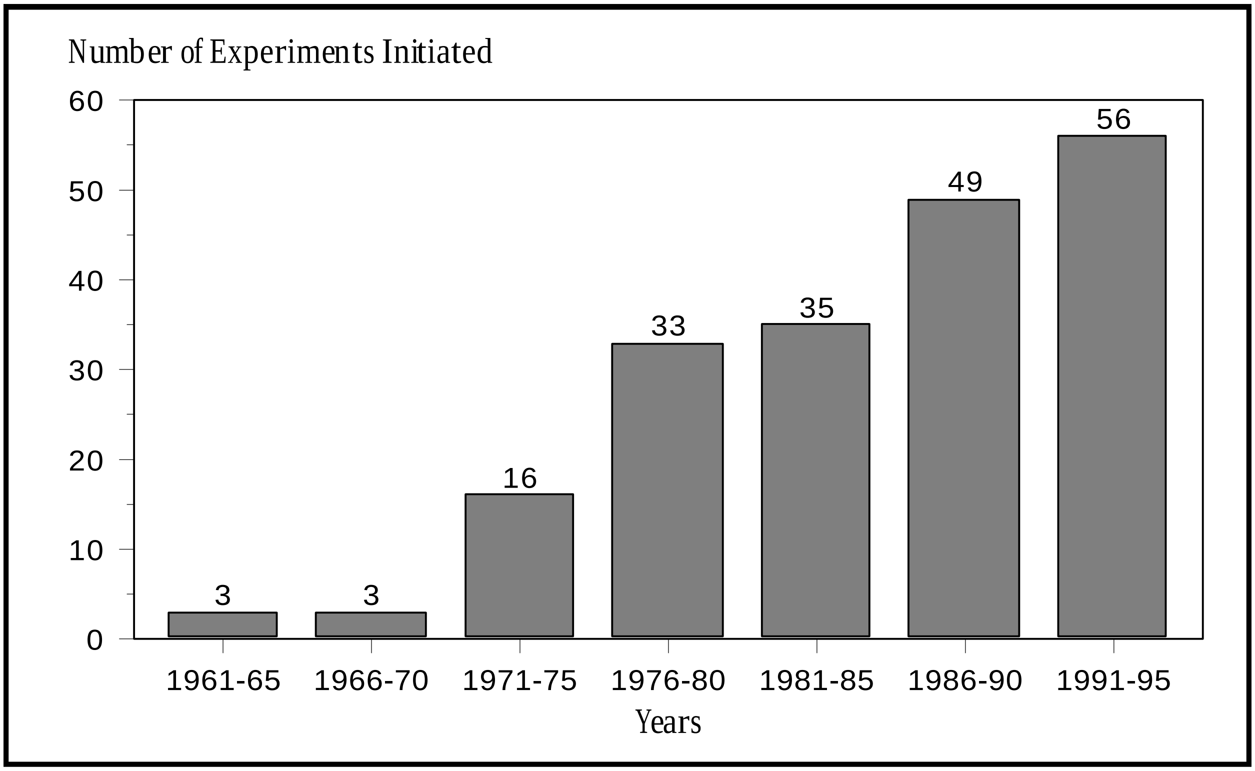
[18] See Botein (1965).

[19] See Riecken and Boruch (1974) for more detailed descriptions of these and other early experiments.

[20] Two other notable developments in research methods were critical to the advent of large-scale field experiments: the development of sophisticated household survey techniques, beginning in the 1940s, and the development of high-speed computers capable of processing large amounts of survey data quickly and efficiently, in the 1950s and 1960s.

Social Experiments Initiated, 1961-1995

EXHIBIT 6



OEO's decision to launch the New Jersey Experiment triggered several more large-scale social experiments. In 1968, in recognition that a large portion of the poverty population resided in rural areas, OEO initiated the Rural Income Maintenance Experiment in Iowa and North Carolina. The following year, the Department of Health, Education, and Welfare (HEW) funded income maintenance experiments in Gary, Indiana, and Seattle, Washington, to test whether the addition of day care subsidies (in Gary) or vocational counseling and/or training (in Seattle) would offset any tendency of cash transfers to cause reductions in work effort. Nearly 5,000 families were randomly assigned in the largest of these four experiments, which ultimately became the Seattle-Denver Income Maintenance Experiment.

These highly visible studies of a politically controversial issue prompted other Federal agencies to adopt this novel technique in the early 1970s. Large-scale experimental tests of housing vouchers, by the Department of Housing and Urban Development, and alternative health insurance plans, by HEW, were explicitly patterned on the income maintenance experiments. In the mid- to late 1970s, a number of experiments were conducted by the U.S. Department of Labor (DOL) and other funding agencies to test alternative employment and training services for unemployed workers, welfare recipients, disadvantaged youths, the mentally impaired, ex-offenders, and substance abusers. Exhibit 6 shows the growth in the number of social experiments initiated over the period 1961–1995. By 1980, 55 social experiments had been initiated; by 1997, 195 experimental studies had begun, according to Greenberg and Shroder (1997).[21]

The growth in acceptance and use of experimental methods to measure the effects of public programs was primarily attributable to two factors: frustration with the failure of nonexperimental methods to yield unequivocal estimates of those effects and the conceptual appeal of the experimental approach. As noted above, a large part of the motivation for the income maintenance experiments was the inability of researchers to obtain consistent estimates of the effects of cash transfers on work effort from existing data. A similar experience with the evaluation of job training programs led to reliance on experimental methods in that area.

In the late 1970s, DOL spent large sums of money on two major evaluation efforts. The first was a series of evalua-

[21] Greenberg and Shroder define "social experiment" as "field studies of social programs in the United States in which there was random assigment of individuals or families to alternative treatments and an emphasis on the measurement of impacts on either market behavior, the receipt of earnings, or transfer payments." Thus, they explicitly do not include in their survey experiments involving interventions such as drug treatment, medical care, or education programs (Greenberg and Shroder, 1997).

tions of its major job training program for disadvantaged workers, the Comprehensive Employment and Training Act (CETA) program. These evaluations were based on data from the Continuous Longitudinal Manpower Survey (CLMS), a follow-up survey of CETA participants, in conjunction with comparison groups drawn from the Current Population Surveys (CPS). During the same period, DOL also funded over 400 demonstrations of employment and training programs for youth under the Youth Employment Demonstration Program Act (YEDPA). Most of these demonstrations had nonexperimental evaluation components.

The CETA evaluations produced widely divergent estimates of the impact of the program on participants' earnings, even though they were all based on essentially the same data.[22] These differences in results were apparently due to differences in the assumptions underlying different nonexperimental methods. And since those assumptions could not be tested or verified, there was no way to know which estimates were most reliable. Moreover, when researchers applied the same set of nonexperimental methods to data drawn from a social experiment, where the experimental estimate provided an unbiased benchmark, they obtained a similar dispersion of estimates.[23] This experience led an expert panel convened to advise DOL on the evaluation of the Job Training Partnership Act (JTPA), the program that succeeded CETA, to recommend strongly that JTPA be evaluated with experimental methods.[24]

Similarly, a National Academy of Sciences committee formed to review the YEDPA demonstrations of the late 1970s concluded that:

> *Despite the magnitude of the resources ostensibly devoted to the objectives of research and demonstration, there is little reliable information on the effectiveness of the programs in solving youth employment problems.… It is evident that if random assignment had been consistently used, much more could have been learned.* (Betsey, Hollister, and Pappageorgiou, 1985)

These recommendations led to the National JTPA Study, in which over 20,000 JTPA applicants in sixteen local programs across the country were randomly assigned either to go into the program or into a control group that was excluded from the program.

On the basis of experiences such as these, a consensus has emerged within the professional evaluation community that random assignment is the method of choice for evaluating public programs. This consensus among the technical experts has led policymakers to accept experimental designs not only as a technical matter, but also as a way to avoid the methodological debates that often accompany the presentation of nonexperimental results, detracting from their credibility and deflecting the policy discussion from substance to method.

Experimental methods are also conceptually appealing to policymakers. In contrast to the arcane statistical sophistication of many nonexperimental methods, the experimental method is relatively simple and intuitively understandable. Even very nontechnical policymakers can appreciate the logic of the experimental contrast between one group exposed to the program and another, which differs from the first only by chance, that is not exposed to the program. This makes experimental studies more accessible and credible to lay persons in the policy process.

For these reasons, not only has the number of social experiments funded and conducted increased enormously over the last two decades, but on a number of occasions, random assignment evaluations have been mandated by Congress. This was the case for the evaluation of the Family Support Act of 1988 and for demonstrations of job training for welfare recipients, self-employment assistance for unemployed workers, and job search assistance for Unemployment Insurance claimants.

Impact of Social Experiments on Policy. Research of any sort is seldom the determining factor in shaping public policy. Experimentation is no exception to this rule. Nevertheless, in part because of their intuitive appeal and credibility, experimental studies have sometimes had decisive effects on policy.

A notable example is the Perry Preschool Project, conducted in the 1960s, in which a sample of 3- and 4-year-old children were randomly assigned either to intensive educational and social services or to a control group that received no special services. A long-term follow-up study of the sample revealed large treatment-control differences in such outcomes as educational attainment at age 19.[25] This study has had a crucial effect on support for intensive early childhood interventions such as Head Start.[26]

[22] See Barnow (1987).

[23] See LaLonde (1986) and Maynard and Fraker (1987).

[24] Stromsdorfer et al (1985).

[25] See Barreuta-Clement et al. (1984).

[26] See Holden (1990).

Another early experiment that had a direct effect on policy was the Manhattan Bail Bond project. Its finding that pretrial release without bail did not increase the incidence of failure to appear for trial led to the incorporation of many of the features of the experimental treatment into the 1966 Bail Reform Act.

The Work-Welfare Experiments, a set of experimental evaluations of state training and job search programs for welfare recipients in the early 1980s, are frequently cited as a major factor in the passage of the Family Support Act of 1988, which established similar programs as national policy.[27] Similarly, the Unemployment Insurance (UI) Self-Employment Demonstrations led directly to national legislation enabling states to provide technical and financial assistance to help unemployed workers become self-employed.[28]

More recently, the results of the National JTPA Study have been influential in decisions by both the Democratic administration and the Republican Congress with respect to funding for JTPA. That study's finding that the program had little or no effect on the earnings of youth was the basis for a substantial reduction in the budget of the youth component and initiation of a systematic search for more effective program models for youth. At the same time, in an era of across-the-board cuts in social programs, funding for the adult component was left intact, largely because the experimental study showed that it was cost-effective.[29]

In these instances, the effects of experimental evaluations have been very clear and direct. More often, such studies have a more subtle influence on the policy process. The income maintenance experiments, for example, added greatly to our knowledge of the labor supply behavior of the low-income population, and therefore conditioned the way income transfer policy was viewed, without leading directly to acceptance or rejection of any specific policy.[30] Similarly, the Health Insurance Experiment produced an enormous amount of valuable information about the relationship between health insurance and the demand for medical services, which has helped inform the national debate on health policy, but was not a decisive factor in the enactment of any specific legislation.

One should not expect that evaluations will determine policy in all cases. Even the best research is only one of many influences in the policy process, and behavioral impacts are only one of many possible legislative objectives. Evaluators will have made a significant contribution to improving the policy process if they provide accurate information on important policy questions, in a form that policymakers can understand. In the remainder of this series of papers, we discuss how experiments can be designed, implemented, and analyzed to achieve that goal.

❦

References

Barnow, Burt S. 1987. "The Impact of CETA Programs on Earnings: A Review of the Literature." *Journal of Human Resources* 22 (Spring): 157-93.

Bell, Stephen H., Larry L. Orr, John D. Blomquist, and Glen G. Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs*. Kalamazoo, Michigan: W. E. Upjohn Institute for Employment Research.

Betsey, Charles L., Robinson G. Hollister, and Mary R. Papageorgiou. 1985. *Youth Employment and Training Programs: The YEDPA Years*. Committee on Youth Employment Programs, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, D.C.: National Academy Press.

Botein, B. 1965. "The Manhattan Bail Project: Its impact in criminology and the criminal law process." *Texas Law Review* 43:319-31.

Burtless, Gary. 1985. "Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment." *Industrial and Labor Relations Review* 39:105-14.

Burtless, Gary, and Larry L. Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources* 21 (Fall): 606-39.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. (1st ed.) Chicago: Rand-McNally.

Fisher, R.A. 1925. *Statistical Methods for Research Workers*. (1st ed.) London: Oliver and Boyd.

Fisher, R.A. 1935. *The Design of Experiments*. London: Oliver and Boyd.

Gramlich, Edward M., and Larry L. Orr. 1975. "The Ethics of Large Scale Social Experimentation," *Ethical and Legal Issues of Social Experimentation*. Washington, D.C.: The Brookings Institution.

Greenberg, David, and Mark Shroder. 1997. *Digest of the Social Experiments*. Washington, D.C.: Urban Institute Press.

Gueron, Judith M., and Edward Pauly. 1991. *From Welfare to Work*. New York: Russell Sage Foundation.

[27] See Gueron and Pauly (1991) for a description of these experiments.

[28] See Orr et al. (1994).

[29] See Orr et al. (1996) for the results of the National JTPA Study and a discussion of its policy impacts.

[30] Some analysts believe that the Seattle-Denver Experiment's finding that cash transfers may lead to marital breakup was responsible for persuading at least one key Senator to withdraw his support of cash transfers to intact families, thereby effectively ending the political prospects for a universal negative income tax.

Heckman, James J., and Robert Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions: An Overview." *Journal of Econometrics* 30:239-67.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (September): 604-20.

Maynard, Rebecca, and Thomas Fraker. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources* 22 (Spring): 194-227

McCall, W.A. 1923. *How to Experiment in Education*. New York: McMillan.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. *Does Job Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington, D.C.: Urban Institute Press.

Orr, Larry L., Stephen A. Wandner, David Lah, and Jacob M. Benus. 1994. *The Use of Evaluation Results in Employment and Training Policy: Two Case Studies*. Paper presented at the Annual Research Conference of the Association for Public Policy Analysis and Management. Bethesda, MD: Abt Associates.

Riecken, Henry W., and Robert F. Boruch (eds.). 1974. *Social Experimentation: A Method for Planning and Evaluating Social Intervention*. New York: Academic Press.

Ross, Heather. 1966. "A Proposal for a Demonstration of New Techniques in Income Maintenance," (mimeo). Data Center Archives, Institute for Research on Poverty, University of Wisconsin, Madison, Wisconsin.

Stromsdorfer, E., H. Bloom, R. Boruch, M. Borus, J. Gueron, A. Gustman, P. Rossi, F. Scheuren, M. Smith, and F. Stafford. 1985. *Recommendations of the Job Training Longitudinal Survey Research Advisory Panel*. Washington, D.C.: Employment and Training Administration, U.S. Department of Labor.

Watts, Harold W., and Albert Rees (eds.). 1977. *The New Jersey Income Maintenance Experiment*. Volume II: Labor-Supply Responses. New York: Academic Press.